学校编码：10384　　分类号＿＿＿＿＿＿密级＿＿＿

学号：200131014　　UDC ＿＿＿＿＿＿＿＿＿＿

学　位　论　文

# 基于小波理论的时序相似性挖掘研究

Similarity Research of Time Series Based on Wavelet in Data Mining

汤　凌　冰

指导教师姓名：罗　键　教授

申请学位级别：硕　　士

专　业　名　称：系　统　工　程

论文提交日期：2004　年 5 月

论文答辩时间：2004　年 6 月

学位授予单位：厦　门　大　学

学位授予日期：

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿＿＿

**2004** 年 **6** 月

# 厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

# 内容摘要

小波分析是当前数学中一个迅速发展的新领域，它同时具有理论深刻与应用广泛的双重意义。与窗口傅里叶变换相比，小波变换是一种灵活的时频局域化变换。通过伸缩和平移等运算功能对函数或信号进行多尺度细化分析，解决了傅里叶变换不能解决的许多难题，被誉为“数学显微镜”，是调和分析发展史上的重要里程碑。

本文试图将小波理论引入股价序列分析，应用于相似性挖掘之中。所谓时序相似性挖掘是指给定一感兴趣时段，运用本文提出的算法从已知时序数据集中提取与之波形相似的时段集。本文拟从以下四方面加以研究：

1. 特征向量的萃取。研究多尺度分析下特征向量与原始数据的逼近问题。使特征向量既具备原始数据的必要表征，又能作一定程度的维数缩减，并给出了基于 DWT 的萃取算法。

2.距离公式的引入。研究欧氏几何空间中的距离模型用以描述相似性，并对其合理性加以证明。

3. 高维索引的构建。研究基于距离的索引结构，给出球面剖分的引理证明并据此提出 VP-TREE 的构建算法。

4. 邻近时序的搜索。研究相邻搜索问题，提出基于 VP-TREE 的搜索算法。

**关键词：**小波变换　特征向量　距离公式　相邻搜索

# ABSTRACT

Wavelet analysis is rapidly developing domain in current mathematic, which is meaningful in both profound theory and extensive application. Comparing with Windowed Fourier Transform, Wavelet Transform has a time-frequency localization of the signal. Trough dilating and translating, it solves a lot of difficult problems by Multi-Resolution Analysis while Fourier Transform can't do .Wavelet analysis is important landmark of harmonic analysis and famous for math microscope.

This paper tries to introduce wavelet theory to stock time series analysis to mine for similarities. The definition of mining time series similarity is picking up time series whose waves are similar to the query time series by algorithm proposed in this article. It will be researched in four aspects as followed：

1．Feature extraction. It begins with researching linear Multi-Resolution approximation between feature vector and the original signal, and then describing how feature vector captures the essence of time series and how the dimension of feature vector is reduced. Finally, it comes out a algorithm based on DWT.

2．Distance formula. The distance model in Euclidean is used for measuring similarity whose rationality is proved.

3．High dimensional index construction. It researches index structure based on distance and proposes algorithm of VP-TREE which is a kind of Ball Partitioning Method.

4．Near neighbor search. It researches the problem of near neighbor search and proposes the search algorithm of VP-TREE.

**keywords**： wavelet transform feature vector distance formula near neighbor search

# 目 录

# TABLE OF CONTENTS

# 第一章 总 论

## 1.1 问题的引入及含义

众所周知，股市运作的随机性虽然较大，但也并非无章可循。实践证明，某些特定波形的出现往往意味着按此趋势买进或卖出，其赢利的机会较大。因而对海量式的股价数据进行有效搜索，从而获得与特定波形有着一定相似性的股价曲线已成为投资决策中一种必不可少的手段。有鉴于此，本文将股价数据纳入视野，加以研究。试图以小波理论为工具，挖掘出相关序列的相似性，从而为投资行为提供强有力的决策支持。要达此目的，需从宏观上弄清两个问题：

1. 如何判断两条股价曲线是相似的？

首先，应对曲线采样，以化连续为离散，使之构成时间序列。显然，一个序列可视为一个向量。故两曲线的相似性在化为两序列的相似性后又转化为两向量的相似性问题。而两向量的相似性可以通过欧氏空间中欧几里德距离加以描述,至此该问题基本获得解决。但问题还可深入下去。尽管时序数据有不同的坐标，尺度与总体趋势，其局部特征却均显现于狭窄的尖峰与拐角，或宽广的山峦与山谷。而小波优良的时频特性使得小波变换能够准确捕捉指定尺度下特定局部的波形特征。所以，运用小波变换将时序数据从时域映射到频域有着天然的合理性。不仅如此，小波变换生成的与原始序列对应的小波系数还可酌情予以筛选，从而减少存储及运算量。经上述处理而成的频域压缩向量，本文谓之特征向量。可以证明，依特征向量相似而获得的结果与依原始时序向量相似所得的结果相比是取伪但不

弃真的。

2. 如何在海量式的时序数据库中进行有效查询？

查询的效率关键在于索引，这一点在海量式数据库中显得尤为重要。如前所述，特征向量不仅能准确反映原始曲线的基本特征，而且长度也比原始序列要小得多，所以索引树应由特征向量来构建。基于球面剖分定理，本文提出了以特征向量为前驱的 VP-TREE 构建与搜索算法，较好的解决了此问题。

## 1.2 课题的研究背景与文献综述

### 1.2.1 基于小波的相似性挖掘

目前，国际上基于 DFT(Discrete Fourier Transform)的时序相似性挖掘算法较为成熟，但由于它不具备灵活的时频局域化与多尺度逼近等优良特性，其主导地位已逐步被 DWT(Discrete Wavelet Transform)所取代。纵览相关文献，小波可从如下三个方面渗入相似性挖掘之中：

- 小波变换可将原始信号从时域转换至频域，并基于一定阈值对小波系数作筛选以实现维数缩减。
- 小波变换可用于抽取特征向量并重新定义相似性标准。
- 小波变换可支持不同尺度下的相似性查询。

小波变换的基本理论可参见文献[16-19;21],小波变换用于时序相似性搜索可参见文献[1；2；3；4]而关于小波方法在时序分析中的应用可参见文献[5；6；7].Chan 和 Fu[3]提出了基于小波的有效时序匹配算法。其中，Haar 小波首次应用于该领域。变换后序列的前几个小波系数充当 R-TREE 索引。通过搜索可获得最近相邻序列。Huhtala et al.[1]则用小波抽取特征向量。Wu et al.[2]将 DFT 与 DWT 在时序相似性挖掘中的应用

作了一个较为全面的比较。Struzik 与 Siebes[8;9]提出了基于 Haar 小波的时序相似性新标准。本文引入小波变换主要用于特征向量的抽取。

### 1.2.2 相似性搜索的有关方法

综观相关文献,不难归纳相似性搜索方法可分类如下:

- Mapping-based on Approaches:
  - Domain-Specific Methods
  - Dimensionality-reduction Methods
  - General Embedding Methods
- Distance-Based on Indexing:
  - Ball partitioning Methods: the VP-TREE; the $VP^{sb}$ - TREE ;
  - the MVP- TREE.
  - Generalized Hyperplane partitioning Methods: the GH- TREE.
  - The M- TREE.
  - The SA- TREE.
  - The distance matrix: ASEA; LAESA.

上述各类方法的详细应用可参阅文献[10-15]。本文仅对 VP- TREE 展开讨论，引入它是为了在海量式的数据库中构建基于特征向量的索引树。另时序数据处理可参见文献[20]。

## 1.3 本文的研究框架

本文从识别股价曲线的相似性出发，先抽象出理论模型予以证明，后提出挖掘算法加以仿真。其研究框架如下所示：

- 第一章是全文的总论。首先给出了问题的引入，然后揭示了问题的

含义。接着阐明了课题的研究背景，并对相关文献作了综述，最后还就本文的研究框架作了说明。其中，研究背景是本文的逻辑起点而文献综述是研究的前提条件。

- 第二章对小波分析及其预备知识作了全面深入的阐述，为进一步研究奠定了坚实的理论基础。
- 第三章对本文提出的挖掘算法进行了准确严谨的描述。既有理论论证，又有算法仿真，是全文的核心所在。
- 第四章与总论首尾呼应，对全文作了总结，并对课题的后续作了展望。

本文的中心在于将 DWT 与 VP-TREE 的构造、搜索算法各取所需，融为一体。

# 第二章　小波分析理论

## 2.1 小波分析的数学基础

本节不加证明的引入一些重要的数学概念，并给出了包含 Hilbert 空间、基以及线性算子在内的一些实分析和复分析结论。

### 2.1.1 函数与积分

模拟信号的数学模型是可测函数。这儿给出 Lebesgue 积分的主要定理。称一个函数 f 是可积的，如果 $\int_{-\infty}^{+\infty}|f(t)|dt<+\infty$。可积函数空间记为 L1(R)。两个函数 f1 与 f2 在 L1(R)中被认为是相等的，如果满足

$$\int_{-\infty}^{+\infty}|f_1(t)-f_2(t)|dt=0.$$

这就是说 $f_1$ 与 $f_2$ 仅仅在一个测度为 0 的点集上不相等即几乎处处相等。

Fatou 引理是关于非负函数列在 Lebesgue 积分下求极限的一个不等式。

**Fatou 引理** 设 $\{f_n\}$ 是一个非负函数族 $f_n\geq 0$，若 $\lim_{n\to+\infty} f_n(t)=f(t)$ 几乎处处成立，则

$$\int_{-\infty}^{+\infty}f(t)dt\leq\lim_{n\to+\infty}\int_{-\infty}^{+\infty}f_n(t)dt.$$

控制收敛定理是函数族存在一个可积函数作为上界的条件下，在 Lebesgue 积分下取极限所得的一个等式。

**控制收敛定理** 设 $\{f_n\}$ 是一个非负函数族 $f_n\geq 0$，若 $\lim_{n\to+\infty} f_n(t)=f(t)$

几乎处处成立。

如果

$$\forall n \in N,\ \left|f_n(t)\right| \le g(t)\ \text{且}\ \int_{-\infty}^{+\infty} g(t)dt < +\infty,$$

则 f 是可积的，且

$$\int_{-\infty}^{+\infty} f(t)dt = \lim_{n\to+\infty} \int_{-\infty}^{+\infty} f_n(t)dt.$$

Fubini 定理给出了交换多重积分顺序的一个充分条件。

Fubini 定理

若 $\int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty}\left|f(x_1,x_2)\right|dx_1\right)dx_2 < +\infty$，则

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x_1,x_2)dx_1dx_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} f(x_1,x_2)dx_1\right)dx_2 = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} f(x_1,x_2)dx_2\right)dx_1$$

### 2.1.2 Banach 空间与 Hilbert 空间

**Banach 空间** 信号常被视作向量。为了定义距离，通常在一个具有范数的向量空间 H 中展开讨论。范数具有如下性质：

$$\forall f \in H, \|f\| \ge 0$$

且

$$\|f\| = 0 \Leftrightarrow f = 0, \tag{2-1}$$

$$\forall \lambda \in C, \|\lambda f\| = |\lambda| \|f\|, \tag{2-2}$$

$$\forall f, g \in H, \|f+g\| \le \|f\| + \|g\|. \tag{2-3}$$

利用此范数，H 中 $\{f_n\}_{n\in N}$ 收敛于 f 是指：

$$\lim_{n\to+\infty} f_n = f \Leftrightarrow \lim_{n\to+\infty} \|f_n - f\| = 0.$$

为保证极限仍在 H 中，需要利用 Cauchy 序列的概念引入一个完备性条

件。序列$\{f_n\}_{n\in N}$称为一个 Cauchy 序列，如果对任意$\varepsilon>0$，当 n 和 p 足够大时，有$\|f_n-f\|<\varepsilon$。若空间 H 中任何一个 Cauchy 序列均收敛于于 H 中一个元素，则称 H 是完备的。

**Hilbert 空间** Hilbert 空间 H 是具有内积的 Banach 空间。两个向量的内积<f,g>关于第一个向量是线性的：

$$\forall\lambda_1,\lambda_2\in C,\langle\lambda_1 f_1+\lambda_2 f_2,g\rangle=\lambda_1\langle f_1,g\rangle+\lambda_2\langle f_2,g\rangle. \tag{2-4}$$

它同时是 Hermitian 对称的：

$$\langle f,g\rangle=\langle g,f\rangle^*.$$

而且

$$\langle f,f\rangle\geq 0$$

且

$$\langle f,f\rangle=0\Leftrightarrow f=0$$

可以验证，$\|f\|=\langle f,f\rangle^{\frac{1}{2}}$是一个范数。非负性蕴含如下的 Cauchy-Schwarz 不等式：

$$|\langle f,g\rangle|\leq\|f\|\|g\|, \tag{2-5}$$

其中等号成立当且仅当 f 和 g 线性相关。

### 2.1.3 Hilbert 空间中的基

**规范正交基** 称Hilbert空间H中元素族$\{e_n\}_{n\in N}$是正交的，如果对$n\neq p$，有

$$\langle e_n,e_p\rangle=0.$$

若对 $f \in H$,总存在序列 $\lambda[n]$ 使得，

$$\lim_{N\to+\infty}\left\| f-\sum_{n=0}^{N}\lambda[n]e_n \right\| = 0,$$

则 $\{e_n\}_{n\in N}$ 称为 H 的一组正交基。由正交性可推出 $\lambda[n]=\langle f,e_n\rangle/\|e_n\|^2$，所以

$$f=\sum_{n=0}^{+\infty}\frac{\langle f,e_n\rangle}{\|e_n\|^2}e_n. \tag{2-6}$$

如果正交基对一切 $n\in N$ 满足 $\|e_n\|=1$，则称之为规范正交基。将式(2-6)两边与 $g\in H$ 取内积就得到规范正交基下的 Parseval 等式：

$$\langle f,g\rangle=\sum_{n=0}^{+\infty}\langle f,e_n\rangle\langle g,e_n\rangle^* \tag{2-7}$$

当 g=f 时，即得到如下的能量守恒率，称为 Plancherel 公式：

$$\|f\|^2=\sum_{n=0}^{+\infty}|\langle f,e_n\rangle|^2. \tag{2-8}$$

**Riesz 基** 在无穷维空间中，如果削弱正交性要求，那么仍然必须加上一个部分的能量等价式以保证基的稳定性。称向量族 $\{e_n\}_{n\in N}$ 是 H 的一个 Riesz 基，如果它是线性无关的，且存在 A>0,B>0 使得对任意的 $f\in H$，总可找到 $\lambda[n]$ 满足：

$$f=\sum_{n=0}^{+\infty}\lambda[n]e_n, \tag{2-9}$$

且

$$\frac{1}{B}\|f\|^2\le\sum_n|\lambda[n]|^2\le\frac{1}{A}\|f\|^2. \tag{2-10}$$

由 Riesz 表示定理可证明存在 $\tilde{e}_n$ 使得 $\lambda[n]=\langle f,\tilde{e}_n\rangle$,并且由(2-10)可推出：